Journal of Computer-Aided Molecular Design, **15:** 41–55, 2001.
KLUWER/ESCOM
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

Comparative molecular field analysis (CoMFA) study of epothilones – tubulin depolymerization inhibitors: Pharmacophore development using 3D QSAR methods

Keun Woo Lee & James M. Briggs*
Department of Biology and Biochemistry, University of Houston, Houston, TX 77204-5513, U.S.A.

Received 3 November 1999; Accepted 20 June 2000

Key words: antitumor agent, CoMFA, epothilone, GFA, 3D QSAR

Summary

A three-dimensional quantitative structure-activity relationship (3D QSAR) study has been carried out on epothilones based on comparative molecular field analyses (CoMFA) using a large data set of epothilone analogs, which are potent inhibitors of tubulin depolymerization. Microtubules, which are polymers of the α/β-tubulin heterodimer, need to dissociate in order to form the mitotic spindle, a structure required for cell division. A rational pharmacophore searching method using 3D QSAR procedures was carried out and the results for the epothilones are described herein. One-hundred and sixty-six epothilone analogs and their depolymerization inhibition properties with tubulin were used as a training set. Over a thousand molecular field energies were generated and applied to generate the descriptors of QSAR equations. Using a genetic function algorithm (GFA) method, combined with a least square approach, multiple QSAR models were considered during the search for pharmacophore elements. Each GFA run resulted in 100 QSAR models, which were ranked according to their lack of fit (LOF) scores, with a total of 40 GFA runs having been performed. The 40 best QSAR equations from each run had adequate fitted correlation coefficients (R from 0.813 to 0.863) and were of sufficient statistical significance (F value from 7.2 to 10.9). The pharmacophore elements for epothilones were studied by investigating the hit frequency of descriptors (i.e. the sampling probabilities of grid points from the GFA studies) from the set of the 4000 top scoring QSAR equations. By comparing the frequency with which each grid point appeared in the QSAR equations, three candidate regions in the epothilones were proposed to be pharmacophore elements. Two of them are completely compatible with the recent model proposed by Ojima et al. [Proc. Natl. Acad. Sci. USA, 96 (1999) 4256], however, one is quite different and is necessary to accurately predict the activities of all 166 epothilone molecules used in our training set. Finally, by visualizing the 35 most probable grid points, it was found that changes related to the C6, C7, C8, C12, S20, and C21 atoms of the epothilones were highly correlated to their activity.

Introduction

Epothilones (Figure 1), which represent a new group of macrocyclic natural products, are widely studied as candidates for a new antitumor agent. Recently, it has been found that the epothilones share their mechanism of action with the structurally dissimilar Taxol$^{\mathrm{TM}}$ (paclitaxel), the well-established antitumor agent [1–3]. Both classes of molecules can bind tightly to the beta-subunit of tubulin and thus inhibit cell division by hyperstabilizing the microtubules such that they cannot dissociate in order to form the mitotic spindle [4–6]. Epothilones are more active, soluble, and available (i.e., from a bacterial brew) than paclitaxel [7]. Additionally, epothilones can be active against paclitaxel-resistant tumor cells [7]. These and other factors increase the potential importance of epothilones in cancer chemotherapy [7]. Epothilones were initially isolated from myxobacteria of the genus *Sorangium cellulosum* [8, 9] and the complete struc-

*To whom correspondence should be addressed. E-mail: jbriggs@uh.edu

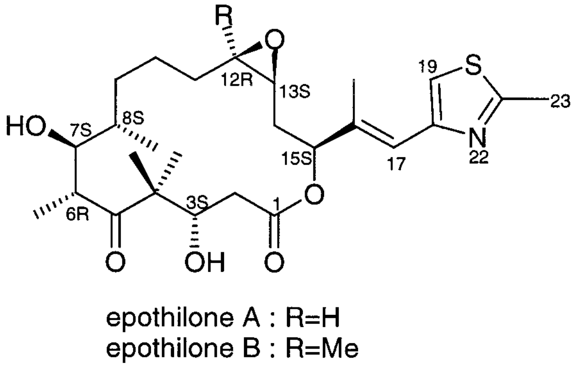


Figure 1. Two-dimensional structure and labeling of epothilone A and B.

ture and stereochemistry were revealed by Höfle et al. in 1996 [10]. The first total syntheses of epothilones were achieved in 1996–1997 [11–15].

The important aims of a QSAR model are to correlate the three-dimensional (3D) structures of drug molecules with their biological activities and to be able to predict the activity of new molecules prior to synthesis [16]. Similarly, the goal of drug design is to be able to design drug candidate molecules with significant efficiency. The information obtained from a pharmacophore model for a lead molecule can play an essential role in drug discovery [16, 17]. At present, one of the most widely used tools for QSAR and drug design is comparative molecular field analysis (CoMFA) [18]. In order to make use of CoMFA, the following four procedures are required: (1) superposition of a set of molecules whose activities have been measured; (2) computation of the interaction energy fields with various probes; (3) statistical analyses to correlate the fields with activities; and (4) interpretation of the coefficients of the resulting QSAR equations. A multivariate data analysis method, e.g., partial least squares (PLS), is usually applied to the CoMFA models in order to derive QSAR equations from the results of descriptor field calculations [19, 20]. PLS is normally used in combination with cross-validation to obtain the optimum number of components. This ensures that the QSAR equations are selected based on their ability to predict the data rather than to fit the data [21]. However, in this work, the genetic function algorithm (GFA) [22, 23] rather than PLS was used to generate multiple QSAR models while searching for pharmacophore elements. Below, we present a CoMFA study on epothilones, which are tubulin depolymerization inhibitors, from which we propose novel pharmacophore elements.

Materials and methods

Biological data and structures

The activity data and 2D structures for 166 epothilone analogs were taken from the literature reported by Nicolaou et al. [24, 25]. The activity data and structure of each molecule are listed in Table 1. Tubulin depolymerization properties were used as activity data. The 3D structures of each molecule were constructed by modification of the X-ray crystal structure of epothilone B [10].

Modeling tools

The computer modeling was performed on a Silicon Graphics, Inc. Octane R10000 computer. Molecular modeling and 3D QSAR studies were performed with the molecular modeling package *Cerius2* (v.4.0) from Molecular Simulations, Inc. (MSI: www.msi.com), including the *QSAR+* module. The *MFA* (*Molecular Field Analysis*) deck in the *QSAR+* module was used for molecular field creation, and the *MSA* (Molecular Shape Analysis) deck for conformational searching and structure alignment, and finally the *GFA* (Genetic Function Algorithm) deck was used for generating the QSAR equation. The manipulation and modification of the molecules were carried out in the *3D builder* module in *Cerius2* and with use of the Merck molecular force field (MMFF) [26, 27] which was used throughout the calculations. The epothilone B crystal structure was compared to the energy minimized structure (using the MMFF forcefield) in order to justify use of that forcefield. The RMSD for all 36 heavy atoms of these two structures was 0.81 Å. The forcefield has been validated against experimental data and compared against all of the major forcefields such as AMBER, CHARMM, CVFF, and OPLS [28].

Conformational searching and alignments

The descriptors and results of the 3D QSAR, CoMFA study are dependent on the 3D structure and hence the conformation of the molecules. Therefore, before constructing QSAR equations, conformations of each molecule in the training set need to be generated, relevant conformations for each molecule chosen, and those conformations aligned with the template structure. The conformational search method used herein is the *Grid Scan* procedure adopted in the *MSA* deck in the *QSAR+* module of *Cerius2*. This method is used to perform a simple systematic search such that

Table 1. Structures and tubulin depolymerization properties of 166 epothilone analogs. The data were taken from the literature reported by Nicolaou et al. (Refs. 24, 25). Bold figures indicate the compounds whose activities are greater than 40%

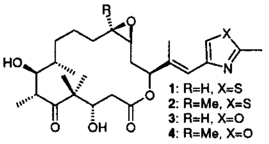
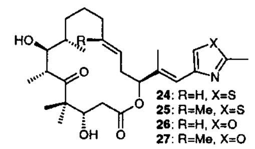
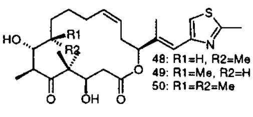
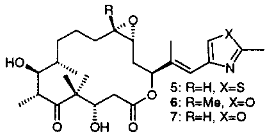
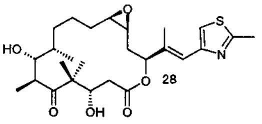
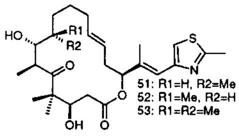
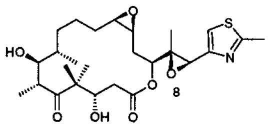
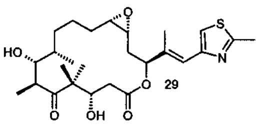
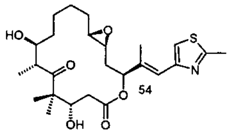
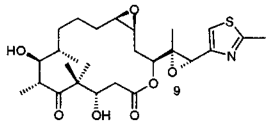
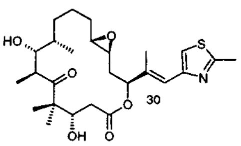
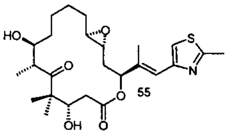
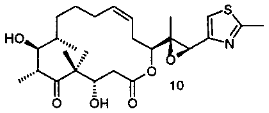
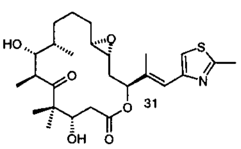
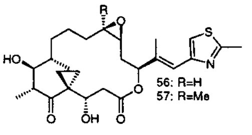
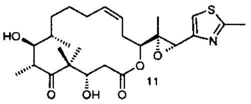
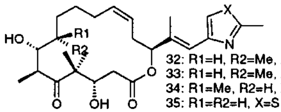
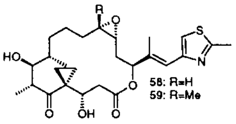
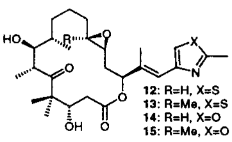
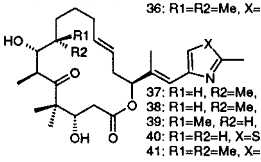
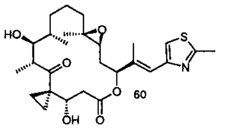
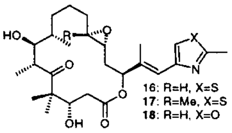
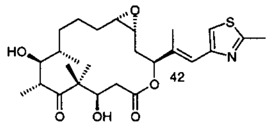
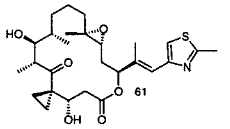
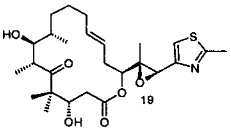
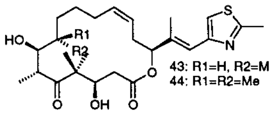
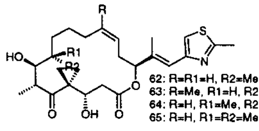
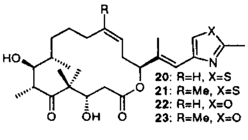
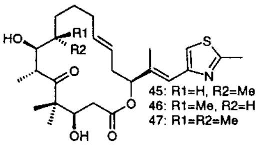
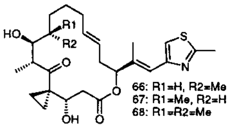
Structure	Tubulin depolymerization (%)	Structure	Tubulin depolymerization (%)	Structure	Tubulin depolymerization (%)
 1: R=H, X=S 2: R=Me, X=S 3: R=H, X=O 4: R=Me, X=O	**76** **98** **58** **93**	 24: R=H, X=S 25: R=Me, X=S 26: R=H, X=O 27: R=Me, X=O	**76** **84** **43** **54**	 48: R1=H, R2=Me 49: R1=Me, R2=H 50: R1=R2=Me	23 5 1
 5: R=H, X=S 6: R=Me, X=O 7: R=H, X=O	17 **71** 24	 28	25	 51: R1=H, R2=Me 52: R1=Me, R2=H 53: R1=R2=Me	24 7 5
 8	16	 29	20	 54	21
 9	14	 30	19	 55	27
 10	20	 31	16	 56: R=H 57: R=Me	20 19
 11	9	 32: R1=H, R2=Me, X=S 33: R1=H, R2=Me, X=O 34: R1=Me, R2=H, X=S 35: R1=R2=H, X=S 36: R1=R2=Me, X=S	13 18 5 23 5	 58: R=H 59: R=Me	5 6
 12: R=H, X=S 13: R=Me, X=S 14: R=H, X=O 15: R=Me, X=O	**92** **84** **64** **95**	 37: R1=H, R2=Me, X=S 38: R1=H, R2=Me, X=O 39: R1=Me, R2=H, X=S 40: R1=R2=H, X=S 41: R1=R2=Me, X=S	13 20 11 21 4	 60	31
 16: R=H, X=S 17: R=Me, X=S 18: R=H, X=O	17 **63** **46**	 42	25	 61	18
 19	12	 43: R1=H, R2=Me 44: R1=R2=Me	16 3	 62: R=R1=H, R2=Me 63: R=Me, R1=H, R2=Me 64: R=H, R1=Me, R2=H 65: R=H, R1=R2=Me	17 22 5 4
 20: R=H, X=S 21: R=Me, X=S 22: R=H, X=O 23: R=Me, X=O	**72** **94** **75** **93**	 45: R1=H, R2=Me 46: R1=Me, R2=H 47: R1=R2=Me	21 3 4	 66: R1=H, R2=Me 67: R1=Me, R2=H 68: R1=R2=Me	25 9 5

Table 1. (continued)

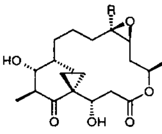
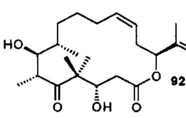
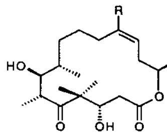
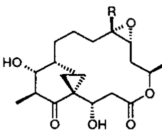
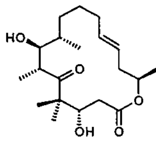
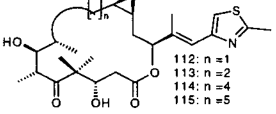
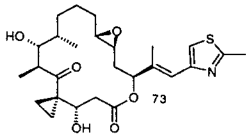
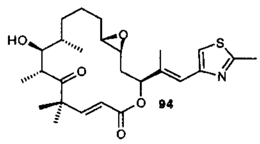
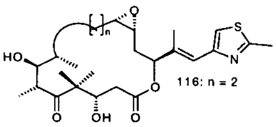
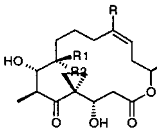
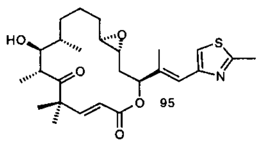
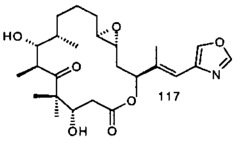
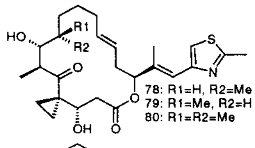
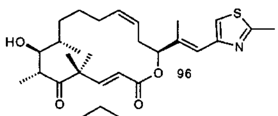
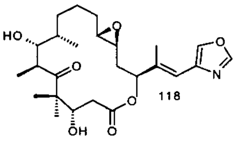
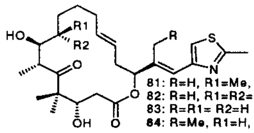
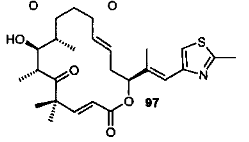
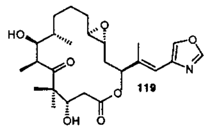
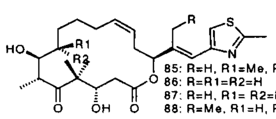
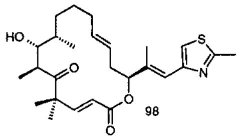
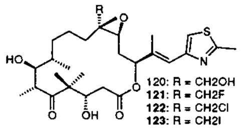
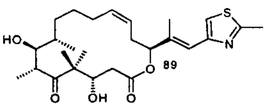
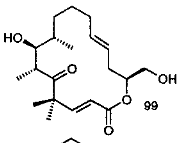
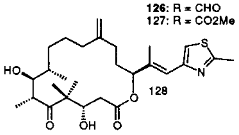
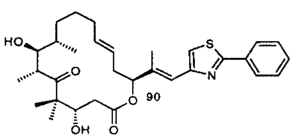
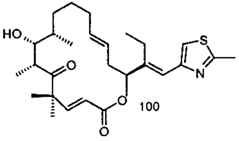
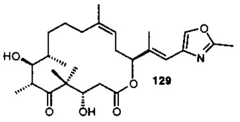
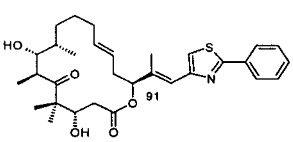
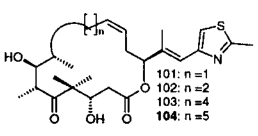
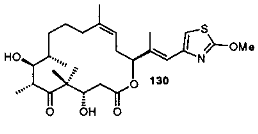
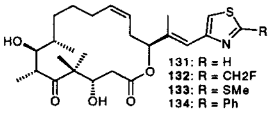
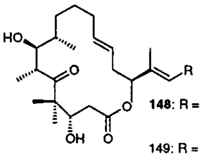
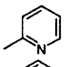
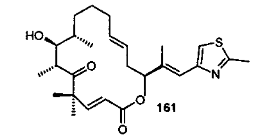
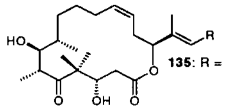
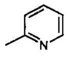
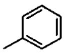
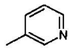
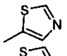
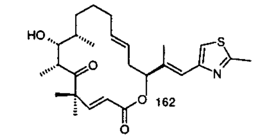
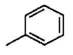
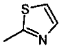
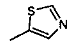
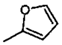
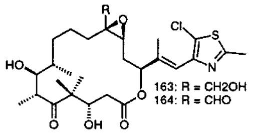
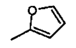
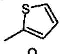
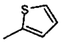
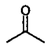
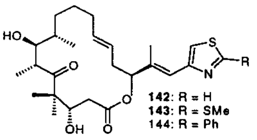
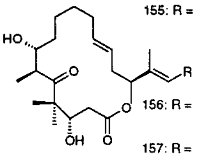
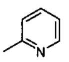
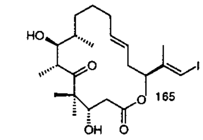
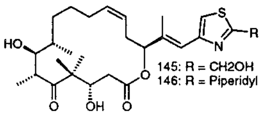
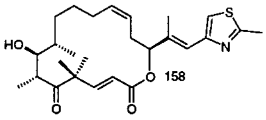
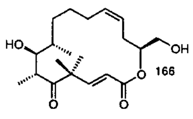
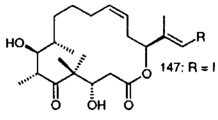
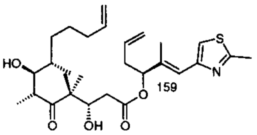
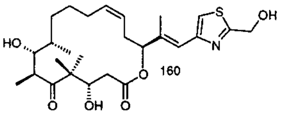
Structure	Tubulin depoly-merization (%)	Structure	Tubulin depoly-merization (%)	Structure	Tubulin depoly-merization (%)
 69: R=H 70: R=Me	10 27	 **92**	**51**	 **105** : R = CHO 106 : R = CO2H **107** : R = CH2Cl **108** : R = CH2F **109** : R = CH=CH2 110 : R = CH2I **111** : R = CH2CH3	**64** 12 **88** **83** **95** 11 **79**
 71: R=H 72: R=Me	9 18	 **93**	**61**	 112: n =1 113: n =2 114: n =4 115: n =5	4 5 7 21
 73	3	 **94**	**46**	 116: n = 2	3
 74: R=R1=H, R2=Me 75: R=Me, R1=H, R2=Me 76: R=H, R1=Me, R2=H 77: R=H, R1=R2=Me	13 6 1 5	 95	28	 117	8
 78: R1=H, R2=Me 79: R1=Me, R2=H 80: R1=R2=Me	18 5 4	 96	27	 118	6
 81: R=H, R1=Me, R2=H 82: R=H, R1=R2=Me 83: R=R1= R2=H **84**: R=Me, R1=H, R2=Me	11 3 24 **48**	 **97**	**58**	 **119**	**71**
 85: R=H, R1=Me, R2=H 86: R=R1=R2=H 87: R=H, R1= R2=Me 88: R=Me, R1=H, R2=Me	11 23 5 34	 98	20	 120: R = CH2OH **121**: R = CH2F **122**: R = CH2Cl **123**: R = CH2I **124**: R = CH2CH3 **125**: R = CH=CH2 **126**: R = CHO 127: R = CO2Me	29 **93** **69** **41** **79** **94** **87** 19
 89	25	 99	17	 128	29
 90	16	 100	39	 **129**	**93**
 91	12	 101: n =1 102: n =2 103: n =4 **104**: n =5	6 9 12 **41**	 **130**	**62**

Table 1. (continued)

Structure	Tubulin depolymerization (%)	Structure	Tubulin depolymerization (%)	Structure	Tubulin depolymerization (%)
 131: R = H 132: R = CH2F 133: R = SMe 134: R = Ph	31 57 92 25	 148: R = 	61	 161	58
 135: R = 	51	149: R = 	26		
136: R = 	13	150: R = 	2	 162	20
137: R = 	16	151: R = 	57		
138: R = 	34	152: R = 	1	 163: R = CH2OH 164: R = CHO	16 5
139: R = 	4	153: R = 	2		
140: R = 	6	154: R = 	1		
141: R =	1	155: R = I	2		
 142: R = H 143: R = SMe 144: R = Ph	41 71 16	 156: R =  157: R = I	8 21	 165	17
 145: R = CH2OH 146: R = Piperidyl	34 18	 158	34	 166	4
 147: R = I	26	 159	18		
		 160	8		

each specified torsion angle is varied over a grid of equally spaced values. While searching for the lowest energy conformer, a cutoff value of 5 kcal/mol was specified relative to the lowest energy conformer. A conformer was accepted if its energy value was less than 5 kcal/mol above the lowest energy conformer.

After obtaining the lowest energy conformers for each molecule, the alignment procedure followed. The global minimum conformation of the most active compound (epothilone B) in the training set was assumed to be the active conformer since the geometry of an epothilone bound to tubulin has not yet been revealed, although a new model has been recently proposed [29]. The active conformer of epothilone B was also used as the shape reference for the remaining 165 molecules. Alignment of structures through a pairwise superposition places all structures in the same frame as the reference compound. The alignment was performed with the MCSG (Maximum Common Sub Group) method adopted in the *QSAR+* module. The method treats molecules as points (i.e., at the atom centers) and lines in three dimensions, and uses the techniques of graph theory to identify patterns. The largest subset of atoms in the reference compound that is shared by all is found and used for the alignment.

Molecular field creation

The molecular field quantifies the interaction energy between a probe molecule and a set of aligned target molecules in a QSAR model. Interaction energies measured and analyzed for a set of 3D structures can be useful in establishing QSARs. To generate an energy field, a probe molecule is placed about a target molecule within a defined 3D grid. At each defined point on the grid, an energy calculation is performed, measuring the interaction energy between the probe and the target molecule. Atoms within the target molecule are not moveable so that the intramolecular energy in the target is ignored since it is a constant. For a set of structures for which energy fields are generated, some of the grid data points can be used as descriptors in generating QSARs and analyzing structure-activity relationships.

For each model, the molecular property field was generated with an OH probe for all molecules of the training set. The OH group, which is a two-atom probe, is introduced as a hydrogen-bond donor and acceptor probe, which consists of an oxygen atom bonded to a hydrogen atom. The van der Waals radii of the atoms were taken from the MMFF forcefield [27]. The charge on the oxygen atom is -0.3881 and that on the hydrogen atom is $+0.3881$, thus the probe has no net charge. Depending on the orientation of this probe, it is capable of behaving as a hydrogen-bond donor or an acceptor. The orientation of this probe is automatically adjusted at each grid point using energy minimization where the oxygen atom of this probe is held fixed at the grid point and only the hydrogen is allowed to move. With this probe, 14-7 van der Waals and Coulomb interaction potentials were calculated to derive the field energy. Each calculation uses a rectangular grid with 2 Å spacing and with 1089 grid points ($11 \times 11 \times 9$). The energy calculation was performed for all grid points such that all energies were constrained to be between -30 and $+30$ kcal/mol. That is, the values which were calculated to be greater than $+30$ kcal/mol or less than -30 kcal/mol were set to be $+30$ or -30 kcal/mol, respectively. Those grid points inside the van der Waals surface are at these extremes. Note that no grid points were initially excluded during the GFA analysis. In other CoMFA procedures, it is common to initially throw out all internal points and all points more than a certain distance away from the van der Waals surface since those points are often highly collinear and not predictive. The GFA method excludes these points during the evolutionary process since they will be found to be non-predictive.

Regression analyses

In order to avoid the covariance or collinearity problem inherent with the multiple linear regression (MLR) method for the systems which use a large number of descriptors, such as from MFA, the partial least squares (PLS) method was used [19–21]. However, the conventional MLR method can produce more significant results than PLS if the number of independent variables (descriptors) is in a reasonable range compared with the number of dependant variables (activities). Generally, the number of compounds should be greater than four or five times the number of descriptors in order to get a reliable model using MLR [30, 31]. If the independent variable sampling method is used, and the number of variables is reasonable, the correlation problem among variables in MLR can be controlled. Especially for CoMFA, the problem can be markedly reduced because a relatively small number of descriptors are used for the model as compared with the huge number of total descriptors (i.e., all grid points). Therefore, with use of a robust sampling method such as a genetic algorithm, one can use the MLR method rather than PLS for evaluating QSARs. The genetic function algorithm was used for descriptor sampling and for the construction of multiple QSAR models. The cross-validation test (leave-one-out) was used for validating the QSAR equations [30, 31].

Genetic function algorithm (GFA)

GFA is a statistical analysis method that generates multiple QSAR models. Usually, this population of models contains many models comparable or superior to the single model generated with standard regression analyses. This method originated from two disparate algorithms: Holland's genetic algorithm and Friedman's multivariate adaptive regression splines (MARS) algorithm [22–24]. Genetic algorithms are derived from an analogy to the evolution of biological systems. An initial population of individuals is randomly created. A fitness function is used to estimate the quality of an individual, so that the best individuals receive the best fitness scores. Individuals with the best scores are more likely to be chosen to mate and propagate their genetic material to offspring through the crossover operation, where pieces of genetic material are taken from each parent and recombined to create the child. After many mating steps, the average fitness of the individuals in the population increases as good combinations of genes are discovered and spread through the population. Genetic algorithms are especially good at searching problem space with a large number of dimensions, as they conduct a very efficient directed sampling of the large space of possibilities. Friedman's MARS algorithm provides an error measure, called the lack of fit (LOF) score, that assigns overall fitness scores based on an evaluation of various features, such as, R^2, F, q^2, bootstrap R^2, etc. [23]. The GFA algorithm uses a genetic algorithm to perform a search over the space of possible QSAR models using the LOF score to estimate the fitness of each model. Such an evolution of a population of randomly constructed models leads to the discovery of highly predictive QSARs.

The analysis was initiated by building a population of 100 randomly constructed equations. This initial population was then evolved up to a given number of generations. The number of populations was chosen such that the values of R^2 and q^2 converged (i.e., 5000 in this study). Evolving the population means that, for each generation, two high-scoring equations are selected as parents and parts of each parent equation are then used to create a child equation. Optional equation mutation operations may be performed on the child at creation. The worst rated equation is then replaced by the new child equation. The multiple linear regression method combined with GFA is used to develop QSAR equations. The GFA search was limited to 35 descriptors, the population size was fixed to 100, and the population was then evolved for 5000 generations, considering the large number of descriptors (derived from 1089 grid points). The number of descriptors was limited to 35 as this is between 4–5 times smaller than the size of the training set, as mentioned above. The mutation probability for adding a new term was given as 50% and only linear polynomial terms were allowed, rather than spline or quadratic ones, for clarity in later interpretation of the QSAR equations. Finally, we obtained the best 100 QSAR equations based on LOF scores for each GFA run. In order to construct multiple QSAR models, 40 independent GFA runs were performed.

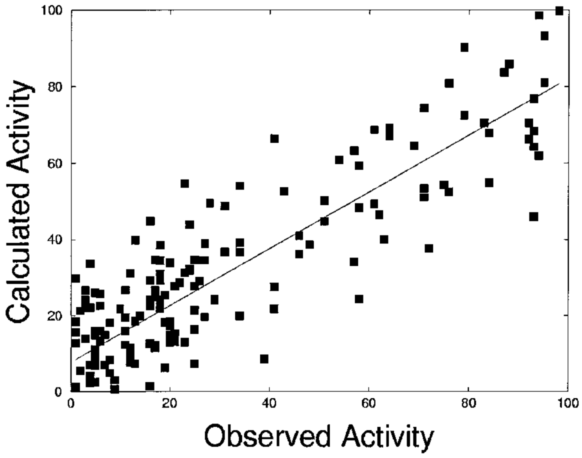


Figure 2. Calculated vs. observed activity plot for the best scored QSAR equation (R = 0.863, F = 10.9) from the 40 equations using 166 compounds in the training set.

Pharmacophore search by investigating the hit frequency of descriptors

The hit frequency of each descriptor (i.e., the sampling probability of each grid point during the GFA runs) in the highest scoring QSAR equations was investigated during the search for pharmacophore elements of epothilones. A pharmacophore element is a region that is highly correlated to the activity of the drug molecules. Analysis of CoMFA models can reveal grid points that are located near the pharmacophore elements and which will then be definitely represented in the QSAR equation. Therefore, we can suggest that the more probable grid points in the best scored QSAR equations are more relevant to biological activity than other points, based on statistical evaluations. In turn, the pharmacophore elements will be near those points. According to the implementation, the most probable points can be determined by counting the hit frequency of grid points in the best-scored QSAR equations.

Each GFA run produced 100 QSAR equations ranked in order of their LOF scores. Thus, 4000 QSAR equations were collected from 40 GFA runs. Each QSAR equation has 35 descriptors which are a subset of the 1089 grid points from the 11 × 11 × 9 lattice grid used in creating the molecular field. Finally, the 140 000 points of grid data (35 grid point descriptors for each of the 4000 QSAR equations) were prepared in order to investigate the hit frequency.

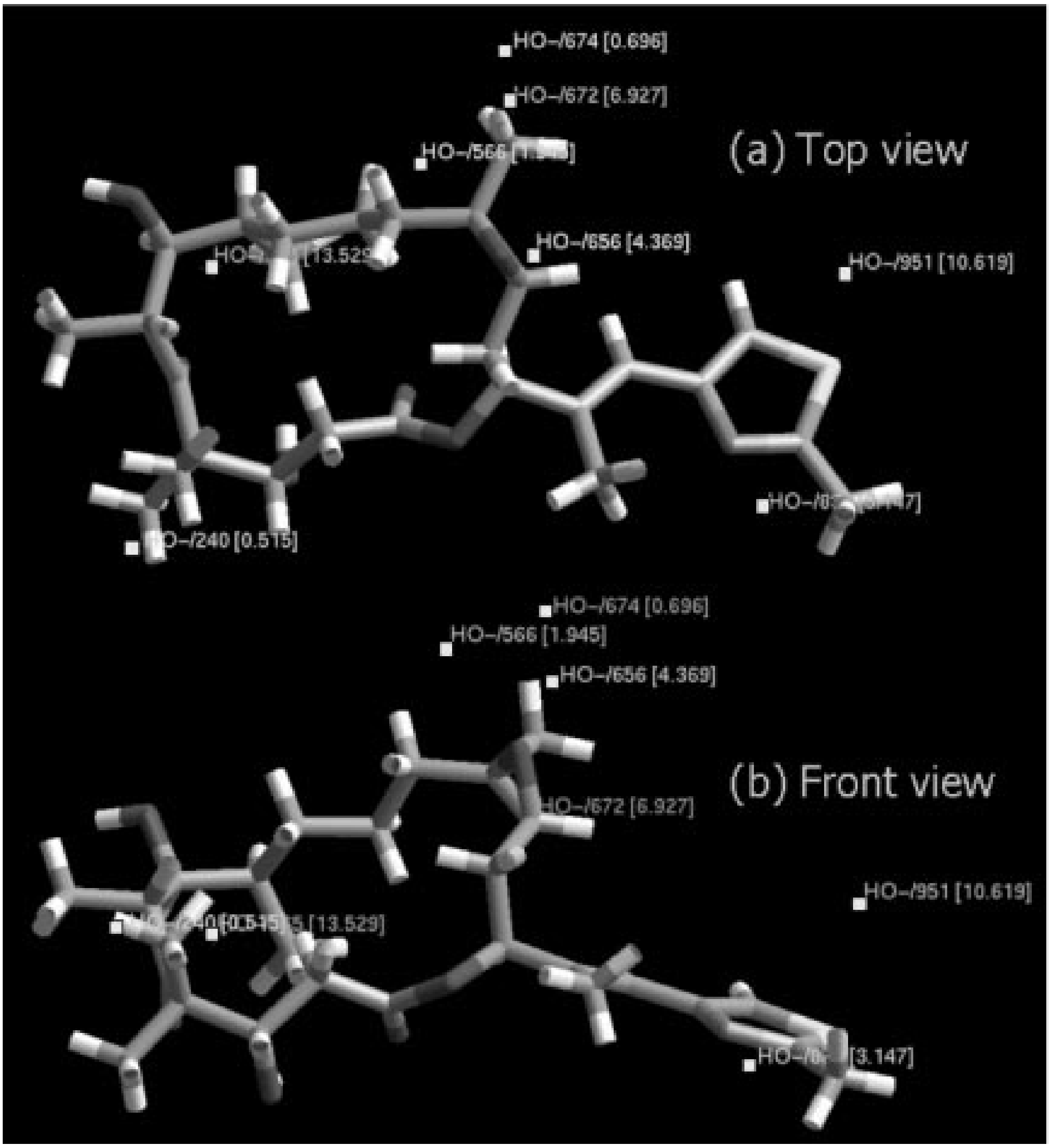


Figure 3. The most probable grid points from the 4000 best scored QSAR equations using 166 molecules and 35 descriptors. The eight most probable grid points near the molecule are displayed as small white boxes. The 3D structure of epothilone B is shown as the reference structure for the 166 epothilone analogs.

Results and discussion

The 166 aligned epothilone analogs were placed in a 11 × 11 × 9 grid lattice. Epothilone B was used as the shape reference for the remaining 165 compounds. Test calculations were performed in order to determine the appropriate grid spacing. A grid spacing of 2 Å was selected as a result of tests with spacings of 1.5, 2.0, and 3.0 Å. The results were better than those using a 3.0 Å grid spacing and very similar to those using 1.5 Å (data not shown). A matrix of 166 × 1089 molecular field energies was generated (i.e., 166 compounds and 1089 grid points). Regression analyses followed,

Table 2. The QSAR results for the 40 best scored equations generated from 40 GFA runs

No.	R^2 [a]	R [b]	F [c]	q^2 [d]	R^2_{BS} [e]	R^2_{BS} error[f]
1	0.739	0.859	10.5	0.558	0.739	0.002
2	0.712	0.844	9.2	0.533	0.712	0.002
3	0.691	0.832	8.3	0.372	0.692	0.002
4	0.681	0.825	7.9	0.408	0.681	0.002
5	0.717	0.847	9.4	0.532	0.717	0.002
6	0.688	0.829	8.2	0.462	0.688	0.002
7	0.702	0.838	8.7	0.511	0.702	0.002
8	0.673	0.820	7.6	0.416	0.673	0.002
9	0.710	0.843	9.1	0.423	0.711	0.002
10	0.678	0.824	7.8	0.490	0.679	0.002
11	0.727	0.853	9.9	0.520	0.727	0.002
12	0.672	0.820	7.6	0.470	0.673	0.002
13	0.669	0.818	7.5	0.459	0.670	0.002
14	0.722	0.849	9.6	0.531	0.722	0.002
15	0.712	0.844	9.2	0.516	0.712	0.002
16	0.712	0.844	9.2	0.516	0.712	0.002
17	0.715	0.845	9.3	0.509	0.715	0.001
18	0.685	0.827	8.1	0.503	0.685	0.002
19	0.663	0.814	7.3	0.443	0.664	0.002
20	0.678	0.823	7.8	0.445	0.678	0.002
21	0.661	0.813	7.2	0.459	0.661	0.002
22	0.684	0.827	8.0	0.437	0.684	0.002
23	0.701	0.837	8.7	0.519	0.701	0.002
24	0.684	0.827	8.1	0.404	0.685	0.002
25	0.679	0.824	7.9	0.460	0.679	0.002
26	0.668	0.818	7.5	0.438	0.669	0.002
27	0.665	0.815	7.4	0.450	0.665	0.002
28	0.696	0.834	8.5	0.469	0.697	0.002
29	0.723	0.850	9.7	0.506	0.724	0.002
30	0.671	0.819	7.6	0.402	0.672	0.002
31	0.719	0.848	9.5	0.431	0.719	0.002
32	0.695	0.834	8.5	0.515	0.696	0.002
33	0.706	0.840	8.9	0.525	0.706	0.002
34	0.675	0.821	7.7	0.444	0.675	0.002
35	0.684	0.827	8.0	0.426	0.685	0.002
36	0.681	0.825	7.9	0.480	0.682	0.002
37	0.704	0.839	8.8	0.459	0.704	0.002
38	0.713	0.845	9.2	0.460	0.714	0.002
39	0.674	0.821	7.7	0.463	0.674	0.002
40	0.745	0.863	10.9	0.619	0.746	0.001

[a] Square of the correlation coefficient.
[b] Correlation coefficient.
[c] F value shows the significance of the equations. Generally, greater than 3.74 means that the confidence limit of the equation is greater than 95%.
[d] Cross-validated $R^2, q^2 = 1 - \mathrm{PRESS}/\Sigma(Y - Y_{\mathrm{mean}})^2$ where PRESS= $\Sigma(Y - Y_{\mathrm{Pred}})$. This value can take values in the range from 1, a perfect model, to less than 0.
[e] Bootstrap R^2, the average squared correlation coefficient calculated during the validation procedure.
[f] Error in bootstrap R^2.

Table 3. The 35 most probable grid points from the 4000 best scored QSAR equations generated by the GFA and MLR methods. The points whose scores are greater than 600 are listed

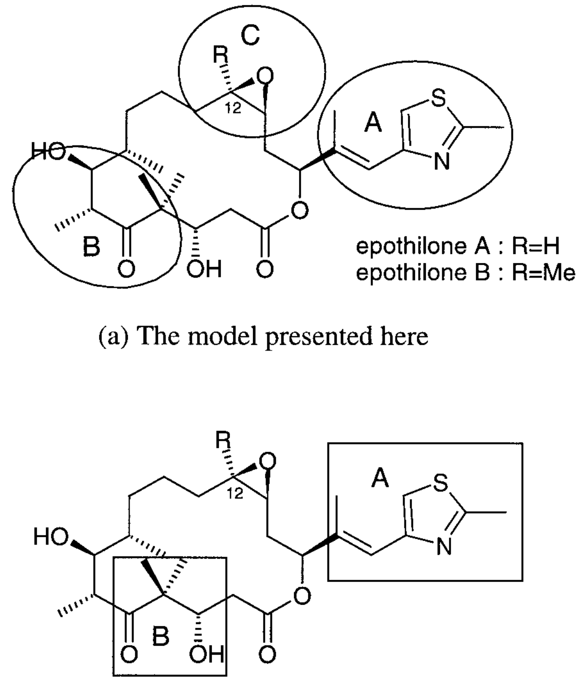
Grid no.	Hit number (Max. 4000)	Grid no.	Hit number (Max. 4000)
28	663	672*	891
69	642	674*	825
172	1700	677	798
181	1101	681	1289
240*	720	682	653
271	931	719	700
279	754	759	1064
365*	3405	764	747
396	928	768	1083
453	2089	777	832
457	605	822	1070
493	677	833*	703
566*	970	860	756
569	620	871	1047
578	831	951*	804
643	911	1059	614
656*	691	1074	600
661	908		

*From these 35 grid points, only eight exist near the molecule and are displayed in Figure 4.

making use of the GFA method and the molecular field energy matrix.

The results of the CoMFA study on epothilones – the statistical analyses for the 40 best scored QSAR equations (one from each GFA run) – are summarized in Table 2. Of the 4000 equations, 40 of them represent reasonable QSAR models for our system. From Table 2, it can be seen that the derived QSAR equations show good correlation coefficients, R (from 0.813 to 0.863), and F values (from 7.2 to 10.9), and therefore indicate considerable correlative capacity and statistical significance of the models. Also, q^2, the cross-validated R^2 values (from 0.372 to 0.619), demonstrate acceptable predictive ability of the models. Using the best-scored QSAR equation (No. 40 in Table 2: $R = 0.863$ and $F = 10.9$), selected from the 40 highest-scoring ones, the relationship between observed and calculated activities was plotted in Figure 2. The figure demonstrates a reasonable positive correlation between the two sets of data.

In order to search for pharmacophore elements of the epothilones, the hit frequency of the descriptors in the 4000 best-scored QSAR equations was investigated. The hit frequencies of the total 140 000 grid



(a) The model presented here

(b) The model suggested by Ojima et al.

Figure 4. Suggested pharmacophore model (a) and Ojima et al.'s model (b) introduced for comparison [32]. Region **A** is definitely the same for both cases, and region **B** is almost the same even though it is a little bit broader in our model. The crucial difference between the two is region **C**. Region **C** can explain many activity changes caused by substituent changes near the C12 atom such as the activity difference between epothilone A and B.

points of data (from 4000 equations × 35 descriptors) were examined. Therefore, the maximum population of each number can range from 4000 to 0. Note that the descriptors (i.e., the relevant grid points) are not the same in all 4000 equations. The results for the 35 most relevant points, where the hits were greater than 600, are listed in Table 3. The higher the hit frequency a grid point has, the more important it is for reliably predicting activity. In order to visualize the relationship between the positions of the grid points and the molecular structure, epothilone B and the most important grid points are displayed in Figure 3 in which only eight points near the molecule are shown. The 3D structure of epothilone B was introduced to show the molecular reference-frame and thus the other structures will deviate from that of epothilone B to some degree. With use of the results displayed in Figure 3, three regions around the epothilones are proposed as representing pharmacophore elements, as depicted in Figure 4a.

Interestingly, while we were preparing this report, Ojima et al. reported on a common pharmacophore model of the Taxol^{TM} family, including the epothilones [32]. Epothilone A and B, elutherobin, discordermolide, and nonantaxel, as well as paclitaxel, were considered. Using NOE data from NMR studies, each active conformation was generated by restrained molecular dynamics simulations and then superimposed for comparison. The structure generated by our conformational search procedure looks similar to that from the Ojima study. Finally, they proposed three regions superimposed on each other as common pharmacophore elements of the Taxol^{TM} family. For comparison with our predictions, the pharmacophore elements proposed by Ojima et al. are given in Figure 4b. Extraordinarily, the epothilones had only two regions superimposed with others (see boxes **A** and **B** in Figure 4b). The comparison shows that region **A** is definitely the same for both the Ojima study and ours. Region **B** is also almost the same in both studies – the crucial difference between the two is the lack of region **C** in the Ojima model. Therefore, unlike our model, the Ojima model cannot explain changes in activity caused by substituent changes near the C12 atom, such as the activity difference between epothilone A and B.

In order to assess the relative importance of each region in the model, a new QSAR equation was constructed using only the 35 most probable grid points as descriptors (Table 3) and the importance of each descriptor for the activity was compared. The linearity (R) and F values for the equation ($R = 0.793$, $F = 19.4$) are compatible with the results of normal GFA calculations (Table 2). By comparing the standardized regression coefficient, the relative importance of each independent variable (descriptor) for the dependent variable (activity) can be estimated. Unlike the regression coefficients which depend on the unit of the variable, the standardized coefficients can be considered as weight factors. Even though they do not reflect the importance in an absolute sense because the values are contingent on the other independent variables, the relative importance of each variable can be assessed. The new equation is summarized in Table 4 along with the results for the 35 points – eight points located near the molecule are shown in bold. These eight points are the most important as they are nearest the molecules. Introduction of appropriate side chains on the corresponding atoms can affect the field energy for those points, and therefore the activity. Our main concern, therefore, was focused on these eight points. The fourth column in the table shows how much each

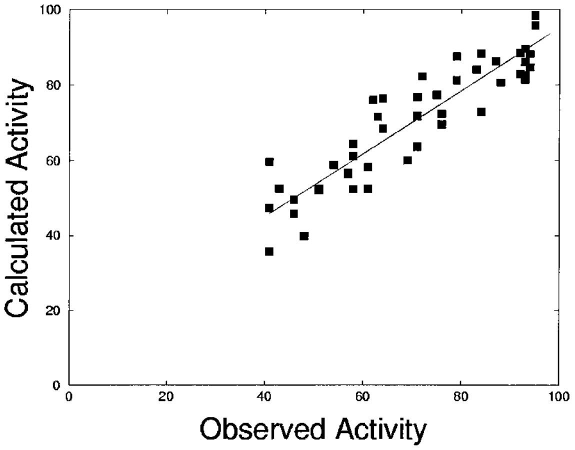


Figure 5. Calculated vs. observed activity plot for the best scored QSAR equation ($R = 0.916$, $F = 18.7$) from the 40 equations using only 47 compounds in the training set that have highest activity (greater than 40% in Table 1).

descriptor contributes to the activity. That is, the absolute value of the standardized regression coefficient indicates the importance of the variable to the activity. For the eight points, the score for point 365 in region **B** is a little bit higher than that for others. This can be interpreted to mean that the energy value at point 365 is more important to the activity than at the other seven. But in region **C**, more points are present. Thus, it can be summarized that the importance to the activity of regions **A** and **B** is similar and that of region **C** is a little bit greater. In view of the molecular structure, it can be concluded that changes related to atoms S20, C21 (region **A**), C6, C7, C8 (region **B**), and C12 (region **C**) were more correlated to the activity of the epothilones than changes made at other atoms.

In order to refine this method, the same CoMFA and hit frequency analyses were performed, although now for only the 47 most active compounds whose activities were greater than 40% (see Table 1). The same calculations were repeated for these compounds, except that the number of descriptors was reduced from 35 to 10 because the number of molecules in the training set was reduced from 166 to 47. The correlation coefficient and the significance for the 40 QSAR models were better (R from 0.863 to 0.916, F from 10.5 to 18.7) than the results with complete training sets (QSAR data not shown). The observed vs. calculated activity plot is given in Figure 5 for the best-scored QSAR equation ($R = 0.916$, $F = 18.7$) of the 40. The results for the hit frequency of the 40 000 grid points of data are not given, however, the five most important grid points near the molecule are shown in Figure 6 for comparison with the results of previous analysis. The overall trend is comparable with the proposed model even though the positions of each point were a little shifted (see Figure 3).

In the pharmacophore search method presented here, introduction of a descriptor sampling method, such as the GFA algorithm, is essential because the determination or sampling of descriptors which will be used in the QSAR equations is not unique for cases with a large number of descriptors, such as CoMFA. That is, it is possible to construct many QSAR equations that use different descriptors, which also have similar linear fit (R) to the activities. Therefore, how can we determine which descriptors are most important? A statistical concept should be applied to overcome the single model problem inherent in GFA. This method can be summarized as the hybrid of the following three concepts: 3D QSAR (CoMFA), Genetic Function Algorithm, and probability theory. QSAR and GFA concepts are used for proper sampling and the hit frequency of each grid point is considered for statistical validation. The following two points are strong features of this method. First, this method does not require the receptor geometry, although this is also true of many QSAR methods. Second, and even more important is that it does not demand the exact bioactive conformation of the molecule. General pharmacophore mapping methods normally require an accurate bioactive conformation. This is due to the conformational search and alignment approach used in this study. If an incorrect conformation is chosen as the template (i.e., not the bioactive conformation), then all other molecules will be aligned such that their equivalent functionalities overlap with the template. Therefore, the search for pharmacophore elements is still valid, although the relative 3D arrangements of the elements may be different, in reality, than predicted. For example, Ojima et al. employed the molecular dynamics (MD) simulation method using NOE data from NMR experiments to generate the active conformation of the molecules studied [32]. However, for finding pharmacophore elements, the present method provides quite similar results without the need for costly MD simulations or NMR data. The results of our method can be used to predict the real pharmacophore structure in the case that the computed global minimum structure is in fact the bioactive conformer. Even when the two structures are different (i.e., the global minimum is not the bioactive conformer), the identification of pharmacophore elements is still very useful for drug

Table 4. The QSAR equation using the 35 most probable grid points as descriptors

No.	Independent variable	Regression coefficient B	Standardized regression coefficient β[b]	Position in the pharmacophore model (see the text)
1	HO-/28	−3.38E+01	−0.391	
2	HO-/69	−5.79E−01	−0.037	
3	HO-/172	6.69E+01	1.077	
4	HO-/181	−9.84E+01	−1.393	
5[a]	**HO-/240**	**1.58E−01**	**0.041**	**Region B**
6	HO-/271	3.06E+01	0.609	
7	HO-/279	2.63E+00	0.094	
8	**HO-/365**	**−1.27E+00**	**−0.295**	**Region B**
9	HO-/396	1.23E+01	0.269	
10	HO-/453	5.38E−01	0.199	
11	HO-/457	1.27E−01	0.074	
12	HO-/493	−5.36E+00	−0.185	
13	**HO-/566**	**3.62E−01**	**0.188**	**Region C**
14	HO-/569	−1.24E+00	−0.050	
15	HO-/578	−4.27E+01	−1.388	
16	HO-/643	8.16E−02	0.033	
17	**HO-/656**	**4.98E−01**	**0.142**	**Region C**
18	HO-/661	3.86E−01	0.184	
19	**HO-/672**	**−2.13E−01**	**−0.124**	**Region C**
20	**HO-/674**	**1.32E+00**	**0.184**	**Region C**
21	HO-/677	4.88E+01	1.448	
22	HO-/681	7.98E−01	0.116	
23	HO-/682	−2.45E+00	−0.138	
24	HO-/719	−8.36E+00	−0.314	
25	HO-/759	−4.29E+00	−0.243	
26	HO-/764	−5.06E+00	−0.308	
27	HO-/768	−1.40E+01	−0.650	
28	HO-/777	4.19E−01	0.015	
29	HO-/822	4.64E−01	0.059	
30	**HO-/833**	**−3.16E−01**	**−0.136**	**Region A**
31	HO-/860	7.15E−01	0.133	
32	HO-/871	5.48E+00	0.268	
33	**HO-/951**	**−4.65E+00**	**−0.191**	**Region A**
34	HO-/1059	−2.15E+01	−0.500	
35	HO-/1074	2.91E+01	0.424	
	intercept	2.82E+01		
	$R(R^2)$	0.793 (0.628)		
	F value	19.4		

[a] The bold figures indicate the eight grid points near the epothilone molecule.

[b] The standardized regression coefficient provides us with a weight factor for the descriptor. Therefore, the magnitude of the absolute value signifies the contribution of the descriptor to the activity. The greater the value, the more important to the activity.

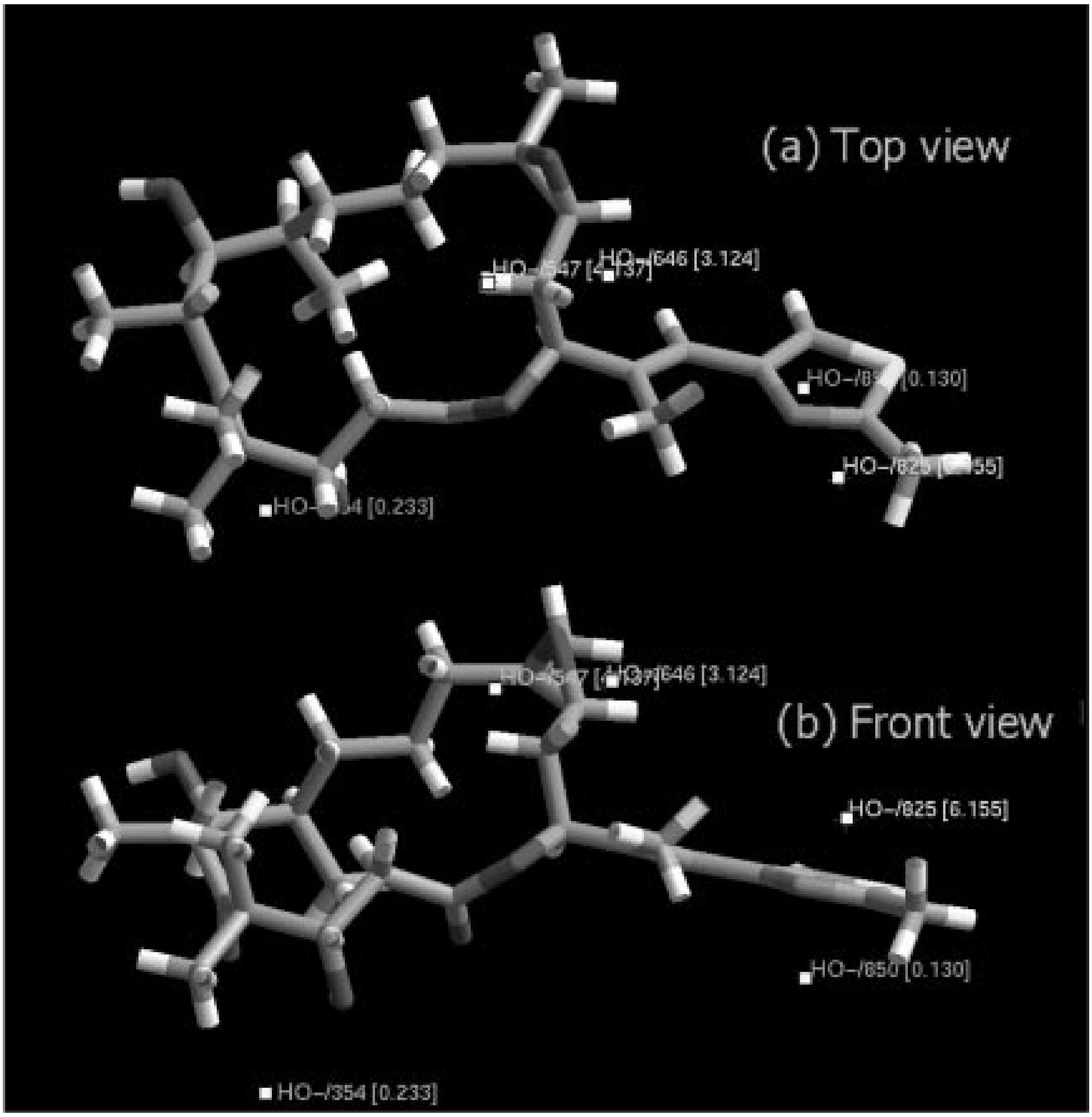


Figure 6. The most probable grid points from the 4000 best scored QSAR equations using the 47 most active molecules from 166, and 10 descriptors out of 1089. The five most probable grid points near the molecule are displayed as small white boxes. The 3D structure of epothilone B is shown as the reference structure for the 47 epothilone analogs.

discovery. Strictly speaking, this method should be regarded as a methodology for active site or pharmacophore element search rather than pharmacophore search because the general systematic conformational search algorithm may not always generate the bioactive conformer. The main flaw of our method is also shared with other QSAR methods, which is that the QSAR equations are only useful if the molecular surface properties are important for activity. For example, if the activity of a series of molecules is largely dependent on, say, dipole moment or transport across a membrane, then a CoMFA-based approach will not be useful.

We have used an OH probe because it can take into account several aspects during a single calculation. It has three characteristics as a molecular field probe, namely, hydrogen-bond donor/acceptor, electrostatic, and steric (vdW). However, it should be noted that this probe resulted in regression coefficients that were difficult to interpret, especially regarding their sign. This is not ideal because it is difficult to decide the direction of change to be made at an important grid point in order to increase activity. On the other hand, this method can address the issue of which grid points are most important. Therefore, using this method with a single character probe, such as a van der Waals only probe or an electrostatic only probe is highly preferred in order to maximize the usefulness of this method. New analyses using single character probes will be pursued in future studies to get more insight into the steric and electrostatic requirements of this system.

In traditional CoMFA studies (i.e., using the CoMFA method as implemented in the Sybyl molecular modeling package from Tripos: www.tripos.com), before analyzing the data, grid points that are more than a certain distance (e.g., 4.0 Å) away from the molecular surface are removed from consideration. Those points cause collinearity problems for PLS since adjacent points far from the molecule tend to have very similar field values. However, this approach was not used in this work because our tools do not have systematic algorithms to accomplish this. Also, we used a GFA method in which bad equations are discarded as the system is evolved through 5000 generations. The final QSAR equations were investigated to look for this problem, however, we could not find any successive and/or meaningless grid points. That is, there were no successive strings of important grid points with energies of ±30 kcal/mol. This means that the GFA procedure has worked well for this method.

A new common pharmacophore model for epothilones and taxanes has just appeared [29]. Using massive conformational searching followed by a tethered energy minimization method, the two classes of compounds were overlaid. The proposed pharmacophore elements, such as the epoxide and the 3-OH groups, are comparable to ours.

Conclusions

The comparative molecular field analysis (CoMFA) methodology was applied to a large data set of epothilone analogs, which are inhibitors of tubulin depolymerization, in order to develop predictive QSAR models. Multiple QSAR models, rather than the usual single model, have been developed using a genetic function algorithm (GFA) to aid in pharmacophore element searching and to address the problem that determining descriptors is not unique in CoMFA. Finally, it was found that the changes related to S20, C21 (region **A**), C6, C7, C8 (region **B**), and C12 atoms (region **C**) were highly correlated to the activity of epothilones. In particular, by considering standardized regression coefficients for the QSAR equation, the importance of regions **A** and **B** was clearly similar but that of region **C** was a little bit greater. A rational pharmacophore search method using the 3D QSAR method was considered and the results for the epothilones were presented. We believe that this idea can be easily applied to other systems during drug discovery.

Acknowledgements

This work was supported by a pilot grant from a Howard Hughes Medical Institute grant awarded to the University of California, San Diego and by a grant from the University of Houston Program to Enhance External Research. Gratitude is expressed to Molecular Simulations, Inc., for providing access to the Cerius2 and InsightII molecular modeling software through the Institute for Molecular Design, at the University of Houston.

References

1. Rwinsky, E.K., Cazene, L.C. and Donehower, R.C., J. Natl. Cancer Inst., 82 (1990) 1247.
2. Ojima, I., Kuduk, S.D. and Chakravarty, S., Adv. Med. Chem., 4 (1998) 69.
3. Ojima, I., Kuduk, S.D., Pera, P. and Veith, J.M., J. Med. Chem., 40 (1997) 279.
4. Shiff, P.B., Fant, J. and Horwitz, S.B., Nature, 277 (1979) 66.
5. Jordan, M.A., Toso, R.J. and Wilson, L., Proc. Natl. Acad. Sci. USA, 90 (1993) 9552.
6. Bollag, D.M., McQueney, P.A., Zhu, J., Hensens, O., Koupal, L., Liesch, J., Goetz, M., Lazarides, E. and Woods, C.M., Cancer Res., 55 (1995) 2325.
7. Kowalski, R.J., Giannakakou, P. and Hamel, E., J. Biol. Chem., 272 (1997) 2534.
8. Höfle, G., Bedorf, B., Gerth, K. and Reichenbach, H. (Gesellschaft für Biotechnologische Forschung, GBR), DE-B 4138042 1993 (Chem. Abstr., 120 (1993) 52841).
9. Gerth, K., Bedorf, N., Höfle, G. and Reichenbach, H., J. Antibiot., 49 (1996) 560.
10. Höfle, G., Bedorf, N., Steinmetz, H., Schomburg, D., Gerth, K. and Reichenbach, H., Angew. Chem. Int. Ed. Engl., 35 (1996) 1567.
11. Balog, A., Meng, D., Kamenecka, T., Bertinato, P., Su, D.-S., Sorensen, E.J. and Danishefsky, S.J., Angew. Chem. Int. Ed. Engl., 35 (1996) 2801.
12. Su, D.-S., Meng, D., Bertinato, P., Kamenecka, T., Balog, A., Sorensen, E.J., Danishefsky, S.J., Zheng, Y.-H., Chou, T.C.,

He, L. and Horwig, S.B., Angew. Chem. Int. Ed. Engl., 36 (1997) 757.

13. Yang, Z., He, Y., Vourloumis, D., Vallberg, H. and Nicolaou, K.C., Angew. Chem. Int. Ed. Engl., 36 (1997) 166.

14. Nicolaou, K.C., Sarabia, F., Ninkovic, S. and Yang, Z., Angew. Chem. Int. Ed. Engl., 36 (1997) 525.

15. Schinzer, D., Limberg, A., Bauer, A., Böhm, O.M. and Cordes, M., Angew. Chem. Int. Ed. Engl., 36 (1997) 523.

16. Martin, Y.C. and Willet, P. (Eds) Designing Bioactive Molecules: Three-Dimensional Techniques and Applications, American Chemical Society, Washington, DC, 1998.

17. Van De Waterbeemd, H., Advanced Computer-Assisted Techniques in Drug Discovery, VCH, Weinheim, New York, Basel, Cambridge, Tokyo, 1995.

18. Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1998) 5959.

19. Wold, S., Ruhe, A., Wold, H. and Dunn III, W.J., SIAM J. Sci. Stat. Comput., 5 (1984) 735.

20. Cramer, R.D., Bunce III, J.D. and Patterson, D.E., Quant. Struct.-Act. Relat., 7 (1988) 18.

21. Kulkarni, S.S. and Kulkarni, V.M., J. Med. Chem., 42 (1999) 373.

22. Holland, J. Adaptation in Artificial and Natural Systems, University of Michigan Press, Ann Arbor, MI, 1975.

23. Friedman, J., Multivariate Adaptive Regression Splines, Technical Report 102, Laboratory for Computational Statistics, Department of Statistics, Stanford University, Stanford, CA, 1988 (revised 1990).

24. Nicolaou, K.C., Vourloumis, D., Li, T., Pastor, J., Winssinger, N., He, Y., Ninkovic, S., Sarabia, F., Vallberg, H., Roschangar, F., King, N.P., Finlay, M.R.V., Giannakakou, P., Verdier-Pinard, P. and Hamel, E., Angew. Chem. Int. Ed. Engl., 36 (1997) 2093.

25. Nicolaou, K.C., Roschangar, F. and Vourloumis, D., Angew. Chem. Int. Ed. Engl., 37 (1998) 2014.

26. Halgren, T.A., J. Am. Chem. Soc., 114 (1992) 7827.

27. Halgren, T.A. and Nachbar, R.B., J. Compnt. Chem., 17 (1996) 587.

28. Halgren, T.A., J. Compnt. Chem., 20 (1999) 730.

29. Giannakakou, P., Gussio, R., Nogales, E., Downing, K.H., Zaharevitz, D., Bollbuck, B., Poy, G., Sackett, D., Nicolaou, K.C. and Fojo, T., Proc. Natl. Acad. Sci. USA, 97 (2000) 2904.

30. Sen, A. and Srivastava, M., Regression Analysis: Theory, Methods and Applications, Springer-Verlag, New York, NY, 1990.

31. Kachigan, S.K., Statistical Analysis: An Interdisciplinary Introduction to Univariate & Multivariate Methods, Radius Press, New York, NY, 1986.

32. Ojima, I., Chakravarty, S., Inoue, T., He, L., Horwitz, S.B., Kuduk, S.D. and Danishefsky, S.J., Proc. Natl. Acad. Sci. USA, 96 (1999) 4256.